This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

On-Request Wireless Charging and Partial Computation Offloading In Multi-Access Edge Computing Systems

Rafia Malik and Mai Vu
Department of Electrical and Computer Engineering, Tufts University, MA, USA
Email: rafia.malik@tufts.edu, mai.vu@tufts.edu

***Abstract*—Wireless charging coupled with computation offloading in edge networks offers a promising solution for realizing power-hungry and computation intensive applications on user-devices. We consider a multi-access edge computing (MEC) system with collocated MEC server and base-station/access point (AP), each equipped with a massive MIMO antenna array, supporting multiple users requesting data computation and wireless charging. The goal is to minimize the energy consumption for computation offloading and maximize the received energy at the user from wireless charging. The proposed solution is a novel two-stage algorithm employing nested descent algorithm, primal-dual subgradient and linear programming techniques to perform data partitioning and time allocation for computation offloading and design the optimal energy beamforming for wireless charging, all within MEC-AP transmit power and latency constraints. Algorithm results show that optimal energy beamforming significantly outperforms other schemes such as isotropic or directed charging without beam power allocation. Compared to binary offloading, data partition in partial offloading leads to lower energy consumption and more charging time, leading to better wireless charging performance. The charged energy over an extended period of multiple time-slots both with and without computation offloading can be substantial. Wireless charging from MEC-AP thus offers a viable untethered approach for supplying energy to user-devices.**

***Index Terms*—Edge computing, wireless charging, energy efficient network, partial data offloading, optimization**

I. Introduction

Multi-Access Edge Computing (MEC) networks have recently garnered significant interest thanks to its ability to provide cloud-computing capabilities within the radio access network, offering proximity, low latency, and high rate access. MEC can bring computing intensive features such as augmented and virtual reality to a large number of connected wireless devices with limited processing capability and battery lifetime by providing services such as computation offloading and wireless charging. Future generation networks offer native support for edge computing functionality, such as key enablers defined by the 3GPP in 5G system architecture to support edge computing [1]. A typical deployment scenario is where the MEC server is co-located with the base-station/access-point (BS/AP) [2]. At the same time, the exponentially growing number of connected devices leads to network densification with a large number of deployed APs. With multiple MEC-APs deployed over a relatively small area in close vicinity to the connected users, RF wireless power transfer from the APs to the user devices becomes practical.

A. Background and Related Work

Computation offloading at the edge has versatile applicability to different use-cases. Examples include (i) AR/VR applications in human-machine interfaces used in smart factories, where complex processing tasks may be offloaded to the edge network, which not only enables easy access to different context information available in the network but also prevents head-mounted AR/VR gear from becoming too warm and uncomfortable to wear [3], (ii) gaming or training service data between two 5G connected devices [4], (iii) real-time map rendering for autonomous vehicular applications [5], and (iv) professional low-latency periodic audio transport services for Audio-Video (AV) production applications, music festivals etc. [6].

Far-field wireless power transfer using Radio Frequency (RF) enables energy-constrained devices to replenish their charge levels without physical connections, offering the inherent advantage of untethered mobility and battery sustainability [7]. There has been significant recent progress in wireless power transfer technology ranging from battery-free cellphone operating on harvested RF energy [8] to reconfigurable RF rectifiers [9]. Commercial products employing RF power transfer have also appeared on the market, charging multiple devices up to 15 meters away [10] [11] [12]. Wireless power transfer in future systems is expected to charge devices at distances ranging from a few meters (for example smart phones) to hundreds of meters (for example sensors) [13]. Adding wireless charging to MEC networks as an *on-request* feature can further help in achieving the required availability and reliability of energy supply, which has become crucial for today's QoS-sensitive applications [14].

Prior works have considered the symbiotic convergence of edge computing and wireless power transfer in different deployment scenarios, for example, wireless charging in cooperation assisted edge computing [15], UAV-enabled mobile edge computing [16] and MEC based heterogeneous networks [17]. Wireless power transfer has been considered in MEC networks for *self-sustained* devices, which rely on wireless charging as their sole power source, in relay-aided edge systems [15], single user [18] and multiple user systems [19]. Such scenarios are typical for devices with low power requirements and/or low receiver sensitivity. Significantly different from this, an *on-request* wireless charging model is where each user-terminal has its own power source and can use wireless charging from the AP to supplement its power consumption. Such *on-request*

charging schemes can minimize the associated energy costs of power transfer and are likely to become an integral part of the maturing 5G vision in the near future [14].

For multiuser edge networks, the transmission strategy and multiple access scheme can significantly impact the overall latency. In terms of communication and data transfer, existing works typically employ sequential protocols like Time Division Multiple Access (TDMA) [20] [19] [18] [21]. Instead, massive MIMO enables simultaneous data offloading from multiple users to the MEC-AP and hence dramatically reduces the wireless transmission time. Employing massive MIMO at the MEC-AP also delivers high throughput and energy efficiency with transmit power savings because of beamforming gains. Massive MIMO can reduce the transmit power at the AP for a given data rate and therefore also has a positive impact on the system energy consumption. In terms of wireless charging, having a large number of antennas at the MEC-AP leads to increased charging range since a larger amount of energy can be reliably directed and transferred [22] [23]. Prior works only consider wireless charging from MEC servers where the AP is equipped with single antenna [20] [15], or having multiple antennas but not with massive MIMO capability [21] [19]. Massive MIMO can be deemed an enabling technology for wireless charging because of its ability to focus energy via sharp beams and charge multiple users concurrently.

B. This Work and Our Contributions

In this work, we consider a multi-cell multi-user network scenario where access points equipped with massive MIMO antenna arrays and with co-located MEC servers offer computation offloading and wireless charging. This model generalizes several existing problems considered in literature on edge computing systems by integrating massive MIMO and power transfer features, which to our knowledge is the first to do so. The computation offloading service is often time critical (for example, due to an upper bound on the motion-to-photon latency for AR/VR applications [2]), and therefore offloading requests by the users must be met within the current time block, leading to the latency constraint. On the other hand, wireless charging is not as time sensitive and a request for charging can be carried out over multiple computation-offloading time blocks. Within each time block, however, the wireless charging occurs at the same time with computation offloading and thus is subject to the same latency constraint.

In each time block, both computation offloading and wireless charging are subjected to the same latency and power constraints. The goal for computation offloading is to minimize the transmitted energy consumption, while the goal for wireless charging is to maximize the amount of received energy. This is different from a joint minimization of energy consumption for both computation offloading and wireless charging, which while consuming less transmitted energy also resulting in a reduced overall received energy. In our formulation, the wireless charging sub-problem is considered secondary and computation offloading sub-problem primary, both are linked by the same power and latency constraints. The wireless charging operation occurs during the MEC-computation phase of the offloading operation. In addition, if computation offloading finishes before the latency limit, the excess time is used for further wireless charging.

The two sub-problems in the considered formulation are not independent but are linked by the same power and latency constraints. Each considered sub-problem is also different from those in the literature. For computation offloading, the energy minimization accounts for energy consumption at both the users and MEC ends, instead of considering only one side [16], [18], [19], [21], [24]. For wireless charging, previous works on active wireless power transfer have no latency constraint, and therefore have a different system model and solution, such as using only the single strongest sub-band for power transfer [25]. The considered wireless charging sub-problem is also different from a self-sustained model which is usually restricted to low-power passive sensors and wearable devices [26]. Here we consider a wireless charging model applicable to an active-user case, such as inside a sports stadium, a conference/exhibition hall, where multiple smart phone users may request wireless charging to replenish battery instead of self-sustaining operations. Here wireless charging is a complementary billable service provided to further enhance the user experience.

Main Contributions

Our main contributions can be summarized as follows.

1) We propose a system model that integrates two MEC services of computation offloading and wireless charging in the same system under the same set of constraints on latency and transmit power. Wireless charging occurs during the MEC-computation phase, and in computation latency-excess time if any. The two sub-problems of computation offloading and wireless charging are treated sequentially, where the objective of computation offloading is to minimize the transmitted energy consumption, and of wireless charging is to maximize the received charged energy. The two sub-problems are coupled together via system latency constraint in each time block.

2) We design a novel, efficient algorithm consisting of two sub-algorithms. The first sub-algorithm optimizes the data partitioning, wireless transmission power and time allocation through a nested-structure using a latency-aware descent algorithm and a primal-dual subgradient algorithm. The derived optimal time allocation is then fed to a second sub-algorithm which finds the optimal energy beamforming matrix (including beam power allocation and beam directions) through a nested structure using a primal dual subgradient algorithm and linear programming.

3) Using our proposed algorithm, we provide detailed quantitative performance analysis and study the impact of different system parameters and optimizing variables on the energy consumption and wireless charging performance. Results show that data partitioning is a key variable affecting system energy consumption, while latency is paramount for wireless charging performance. The optimal charging beams can also use the sidelobes and backscatter as a means of increasing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

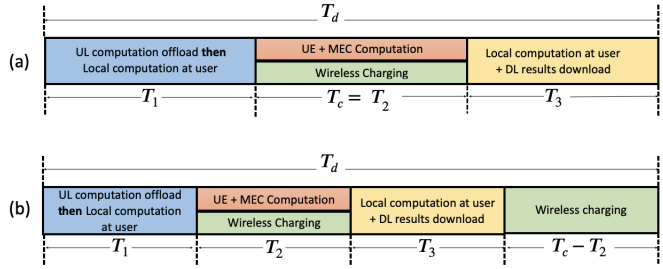

Figure 1: Timing diagram and functional model of the system's operation for both data computation offloading and wireless charging. Wireless charging is performed during MEC computation period when latency is tight (a), or both during MEC computation and after computation offloading finishes when latency is not tight (b).

their harvested energy over the latency constrained charging time. Successive wireless charging over an extended period of multiple time blocks, both with and without computation offloading, can deliver a significant amount of energy to the users.

Notation: $\boldsymbol{X}$ and $\boldsymbol{x}$ denote a matrix and vector respectively, $\nabla^2 f(x)$ denotes the Hessian, and $\nabla^2 f(x)^{-1}$ denotes its inverse. For an arbitrary size matrix, $\boldsymbol{Y}$, $\boldsymbol{Y}^*$ denotes the Hermitian transpose, and $\mathbf{diag}(y_1, ..., y_N)$ denotes an $N \times N$ diagonal matrix. $\boldsymbol{I}$ denotes an identity matrix, and $\mathbf{0}, \mathbf{1}$ denote an all zeros and all ones vector respectively. The standard circularly symmetric complex Gaussian distribution is denoted by $\mathcal{CN}(\mathbf{0}, \boldsymbol{I})$, with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$. $\mathbb{C}^{k\times l}$ and $\mathbb{R}^{k\times l}$ denote the space of $k \times l$ matrices with complex and real entries, respectively.

II. System Model

We consider a system where $L \geq 1$ Access Points (APs), each co-located with an MEC Server, are deployed over a targeted zone/area, for instance in a sports stadium, town fair or a conference exhibition hall, serving ground users with computation offloading and power transfer. Each AP is equipped with a massive antenna array with N antennas while the user-devices are equipped with single antennas. These APs wirelessly charge (upon request) ground users in downlink, collect offloaded data from the users in uplink, and deliver computed results to users in downlink [27]. We consider K users requesting wireless charging service and sending data for computation offloading to each MEC-AP. In the case of cellular networks, wireless charging can be a billable service assuming that the ground users have knowledge of their battery state, and can request the AP for wireless recharging when their battery is critically low.

Consider the case where wireless charging is requested jointly with computation offloading, which includes the scenario of charging only or computation only as special cases. There are three functions contributing to the system's operation as shown in Figure 1; (i) wireless charging of the user terminals by the MEC-AP, (ii) data transmission in the form of computation offloading from the users to the MEC-AP in the uplink and results downloading from the MEC-AP to the users in the downlink, and (iii) data computation at the MEC server and locally at the users.

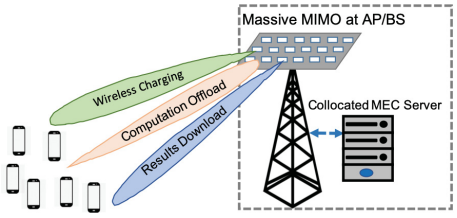

Figure 2: Conceptual beamforming structure for maximal information and energy transfer using a massive MIMO antenna array at the MEC-AP

Given a latency constraint of T_d, the time span for data offloading, computation at both the users and the MEC ends, wireless charging, and delivery of computed results to the user must not exceed T_d. Considering computation offloading, this operation is divided into three timing phases: The maximum time duration for data offloading to the MEC is denoted by T_1, the computation for offloaded data at the MEC spans duration T_2, and the transmission of processed results occupies time T_3. The timing during for wireless charging will be dependent on these three computation offloading phases and the total latency. Figure 1 shows two scenarios timing model: either computation offloading requires the whole duration of T_d, in which case the wireless charging is restricted to the computation phase, or computation offloading consumes a time duration less than T_d and therefore wireless charging can continue after computed results have been transmitted in downlink. Our formulations in the next section account for both of these scenarios. We discuss the energy and time consumption of each system's function, namely wireless charging, data transmission and data computation.

It is worth noting that while it is possible for the user to offload data and perform local computation at the same time during Phase I, the user's power is limited, hence we assume that the i^{th} user focuses its power for offloading data and performs no computation during $t_{u,i}$, the time during which it is offloading its own data. However, the i^{th} user may perform local computation during Phase I as soon as the offloading is done, that is after $t_{u,i}$ for $t_{u,i} < T_1$, and may continue local computation, if required, during Phase II and Phase III since the time for offloading $t_{u,i}$ and local computation $t_{L,i}$ for a user is also bounded by the latency constraint, that is, $t_{u,i} + t_{L,i} <= T_d$. This condition is specified later in constraint (13d) in our problem formulation. Other works in multi-user computation offloading consider local computation even during the current user's offloading time, however, such problems are based on sequential TDMA offloading in which each user is slotted a separate and sequential time duration for offloading, without considering the time spent in phases II and III [19]. Our work, on the other hand, considers simultaneous offloading from all users and also accounts for the time spent in phases II and III in the overall optimization. Based on our previous results which show that the offloading time in uplink is typically much smaller than the local computation time for a user [5], we expect changes in total time or energy consumption to be insignificant even if we allow local computation during the offloading time.

A. Wireless Charging

In each cell, we consider K users requesting wireless charging from the MEC-AP, where the i^{th} user requests e_i mJ of energy. To cater for the energy requests from multiple users, the massive-MIMO enabled MEC-AP employs transmit energy beamforming, as shown in Figure 2. Such energy beamforming requires channel state information (CSI), which can be obtained at the AP using *uplink training*, where pilot symbols are transmitted over some duration of the coherence interval to estimate the channel matrix from the users to their serving MEC-AP. For the downlink channel, we assume Time Division Duplex (TDD) operation such that the channel matrix from the AP to the users can be obtained by wireless channel reciprocity of the uplink channel and hence the transmission of downlink pilots becomes unnecessary [28] [29].

Let $\boldsymbol{x_q}$ denote the energy bearing signal from the AP to the user-terminal (UT), $\boldsymbol{W_q} \triangleq \mathbb{E}\left[\|\boldsymbol{x_q}\|^2\right]$ denote the transmit covariance matrix, and $P_c = \text{tr}(\boldsymbol{W_q})$ be the power transmitted from the AP for wireless charging, in short, the charging power. Then the received (charged) power at the i^{th} user is given as

$$P_{h,i} = \xi_i \mathbb{E}\left[\left|\boldsymbol{h_i^* x_q}\right|^2\right] = \xi_i \text{tr}(\boldsymbol{h_i^* W_q h_i}) \tag{1}$$

where $0 \leq \xi_i \leq 1$ is the energy conversion efficiency from Radio Frequency (RF) to Direct Current (DC) for the i^{th} user and $\boldsymbol{h_i} \in \mathbb{C}^{N\times 1}$ is the channel from the AP to the i^{th} user. We assume a linear energy harvesting model where the energy conversion efficiency per user is constant over a single time block duration, T_d. Non-linear wireless charging models, with a variable energy conversion efficiency over time, are more applicable to scenarios where there is a variation in the received power [30] such as at high SNR and also depends on the rectifier characteristics (diode breakdown region) [31]. For the considered system with strict latency constraint on each time block for on-request charging, constant user energy conversion efficiency is more suitable. To account for the difference in received power at each user location, each i^{th} user has its own energy conversion efficiency ξ_i based on the received power in the current time block.

We define T_c as the time duration for wireless charging, where $T_c = T - (T_1 + T_3)$ and includes the time consumed by the computation phase, over which power is transferred to the users alongside computation at the users and at the MEC server. The energy consumed at the MEC server for power transfer, in short the charging energy, is given by

$$E_c = T_c \text{tr}(\boldsymbol{W_q}) \tag{2}$$

We consider a *wireless charging maximization* approach where the received energy at the users is maximized subject to the latency and MEC-AP's transmit power constraint. For the i^{th} user requesting e_i amount of energy, the received (charged) energy, $E_{h,i}$, is constrained as below

$$E_{h,i} = P_{h,i} T_c = \xi_i T_c \text{tr}(\boldsymbol{h_i^* W_q h_i}) \leq e_i \;\forall\; i \in [1, K] \tag{3}$$

Here the amount of wireless charging is upper-bounded by e_i such that the charged energy is at most equal to the requested amount so as not to overcharge the users since charging is a billable service, and also not to burn the user's battery. Having this bound ensures feasibility of energy transfer. In this way no single user gets an unfairly large amount of the charged energy at the expense of others, and only a portion of the requested energy may be charged (in the current time block) if it is unfeasible for the AP to satisfy the user's energy request completely due to poor channel conditions or high energy request(s) by a single or few users.

Note that in cases where the i^{th} user's energy request is only partially fulfilled in the current time block, the remaining amount may be charged in subsequent time blocks. Since, in our considered system model, the operation of computation offloading is not dependent on wireless charging for energy, charging can be deferred to future time blocks if computation offloading demands more time and energy resources in the current time block. For the remaining of the paper, to simplify notation, we assume that for the current time block, no amount of energy has previously been received by the user, and the charge requested by the i^{th} user is equal to e_i. All subsequent formulations and algorithms, however, are applicable if this requested energy is scaled by a factor to reflect a proportion in each time block.

B. Data Transmissions

For computation offloading at each MEC, we consider the simple *data-partition model*, where the task-input bits are bit-wise independent and can therefore be arbitrarily divided into different groups to be executed by different entities [32]. We consider the case of partial offloading, such that for the i^{th} user, the u_i computation bits are partitioned into q_i and s_i bits, where q_i bits are computed locally and s_i bits are offloaded to the MEC server. Assuming that such partition at the user-terminal does not incur additional computation bits, then $u_i = q_i + s_i$.

1) Offloading Data in Uplink: In a given time slot, K single-antenna user terminals simultaneously offload to the N antenna AP. We consider $N \gg K$ such that the throughput becomes independent of the small-scale fading with channel hardening [29]. The very large signal vector dimension at a massive MIMO AP enables the use of linear detectors such as maximum ratio combining (MRC), in which case the uplink net achievable transmission rate for the i^{th} user in the l^{th} cell, $r_{u,i}$, is given as [28]

$$r_{u,i} = \nu \log_2\left(1 + \frac{\text{SINR}_{li}^{ul}}{\Gamma_1}\right),\; \text{SINR}_{li}^{ul} = \frac{N\gamma_{li}^{l} p_{li}}{\sigma_{1,li}^2} \tag{4}$$

where $\Gamma_1 \geq 1$ accounts for the capacity gap due to practical coding schemes, γ_{li} is the mean-square channel estimate, and p_{li} is the transmit power of the i^{th} user in the l^{th} cell. The constant ν represents the portion of transmission symbols spent on data transfer in the coherence interval τ_c. The interference and noise power, $\sigma_{1,li}^2$, includes the receiver noise variance, interference due to channel estimation and

from contaminating cells, and inter-cell interference as defined in [28, Eq. 4.18], and is dependent on all users' transmit power and channel conditions [5].

The energy consumed for offloading the i^{th} user's data is given by $E_{OFF,i} = p_i t_{u,i}$, where p_i is the transmit power and $t_{u,i}$ is the transmission time for the i^{th} user. Let B denote the channel bandwidth, then $t_{u,i} = \frac{s_i}{Br_{u,i}}$. All users offload their computation bits simultaneously, and the total energy and time overhead for simultaneous data offloading is given as

$$E_{OFF} = \sum_{i=1}^{K} \frac{p_i s_i}{Br_{u,i}}, \; T_1 = \max_{i\in[1,K]} t_{u,i}. \tag{5}$$

2) Downloading Results in Downlink: For the i^{th} user in the l^{th} cell, the downlink transmission rate with maximum ratio linear precoding at the MEC-AP is given as [28]

$$r_{d,i} = \log_2\left(1 + \frac{\text{SINR}_{li}^{dl}}{\Gamma_2}\right), \; \text{SINR}_{li}^{dl} = \frac{NP\gamma_{li}^{l}\eta_{lk}}{\sigma_{2,li}^2} \tag{6}$$

where $\Gamma_2 \geq 1$ is the capacity gap, and $\sigma_{2,li}^2$ is the interference and noise power which also contains pilot contamination and intercell interference as given in [28, Eq. 4.34], and depends on the power allocation at the MEC-AP for downlink wireless transmission and also on the channels between the AP and the users [5].

The transmission time for delivering the i^{th} user's computation results can be written in terms of the downlink rate in (6) as $t_{d,i} = \frac{\tilde{s}_i}{Br_{d,i}}$. Here $\tilde{s}_i$ denotes the number of information bits generated after processing s_i offloaded bits of the i^{th} user. The number of information bits generated as a result of data computation ($\tilde{s}_i$) are proportional to the data bits to be computed (s_i), that is $\tilde{s}_i \propto s_i \rightarrow \tilde{s}_i = \mu s_i$. μ is the proportionality parameter between the amounts of requested and computed data and is not restricted to the range [0,1], rather it adds an application-centric flexibility to our system model in terms of the data size in downlink. For instance, $\mu < 1$ for face recognition applications or $\mu \gg 1$ for video-rendering applications [33] [34] [5]. The AP simultaneously transmits computed results for all users, and the total energy and time overhead for results downloading are then given as

$$E_{DL} = \sum_{i=1}^{K} \frac{P\eta_i \mu s_i}{Br_{d,i}}, \; T_3 = \max_{i\in[1,K]} t_{d,i}. \tag{7}$$

C. Data Computation

1) Local computation at the users: The time for computation depends on the amount of data to be computed and the CPU cycle frequency. The energy consumption and the processing time for local computation at the i^{th} user is given as [32]

$$E_{LC} = \sum_{i=1}^{K} \kappa_i c_i (u_i - s_i) f_{u,i}^2, \quad t_{L,i} = \frac{c_i(u_i - s_i)}{f_{u,i}} \tag{8}$$

where κ_i is the effective switched capacitance, $f_{u,i}$ denotes the average CPU frequency, c_i denotes the CPU cycle information, and $q_i = u_i - s_i$ is the total number of bits required to be locally computed at i^{th} user respectively.

2) Computation of the offloaded data at the MEC server: MEC servers, with high computation capacities, compute the tasks of all users in parallel [35] [32]. The energy and time consumed for computing offloaded bits is given as

$$E_{OC} = \sum_{i=1}^{K} \kappa_m f_{mi}^2 d_m s_i,$$
$$t_{M,i} = \frac{d_m s_i}{f_{mi}} \; \forall i \in [1,K], \quad T_2 = \max\{t_{M,i}\} \tag{9}$$

where $t_{M,i}$ is the time for computing i^{th} user's offloaded task, s_i is the number of bits offloaded by the i^{th} user to the MEC, d_m is the number of CPU cycles required to compute one bit at the MEC, f_{mi} is the CPU frequency assigned to the i^{th} user's task, and κ_m is the effective switched capacitance of the MEC server. The computation at the MEC is synchronous such that computation only begins after data from all users has been offloaded. While it is possible to perform fine-scale timing optimization where the MEC starts computing immediately after it receives a user's data, the expected gain from this would be negligible since the computation time, T_2, is short compared to T_1 and T_3 [5, Figure 6] and further optimizing each user's computation time at the MEC can significantly increase the formulation and algorithm complexity.

Previous results in [5] show that in a typical network setting, the wireless transmission energy consumption is significantly dominant compared to the computation energy consumption. Only when a user is located within a few meters from an MEC-AP, the wireless transmission energy may be reduced substantially to be comparable with computation energy. Therefore, for our formulation to follow in Section III, we consider equal frequency allocation for users' tasks, that is $f_{m,i} = f_m \; \forall i$ since dynamic frequency allocation has little effect on the overall system's energy consumption compared to wireless transmission.

III. Optimization Problem Formulations

Considering a multi-cell multi-MEC network, we formulate an edge computing problem which explicitly accounts for physical layer parameters including available transmit powers from each user and the MEC, associated massive MIMO data rates with realistic pilot contamination and interference. For simplicity of notation, we assume that all K users which are offloading their computation to the MEC server are also requesting wireless charging.

In this section, we discuss a sequential formulation and consider the problems of computation offloading (P_{CO}) and wireless charging (P_{WC}) independently in terms of energy optimization. The aim of (P_{CO}) is to minimize the energy consumption for computation offloading, while the goal of (P_{WC}) is to maximize the energy received at the users through wireless charging. Wireless charging happens during the time available after timing has been optimally allocated for computation offloading. This leads to a sequential optimization process where the optimization for wireless charging will follow that of computation offloading. It should be emphasized

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

that the sequential process is only in terms of optimization, as once all the variables and parameters are optimized, the operations of computation offloading and wireless charging can occur simultaneously as discussed in the system model of Section II.

A. Minimization Of Energy Consumption For Computation Offloading

Using the uplink and downlink transmission rates, respectively defined as $r_{u,i} = \frac{s_i}{\nu t_{u,i} B}$ and $r_{d,i} = \frac{\mu s_i}{t_{d,i} B}$, and based on (4) and (6), we can express the per-user power allocation variables for uplink (p_{li}) and downlink (η_{li}) transmissions as functions of the time allocation and data partitioning as follows:

$$p_{li} = \frac{(2^{\frac{s_i}{\nu t_{u,i} B}} - 1)\Gamma_1 \sigma_{1,i}^2}{N\gamma_i}, \quad \eta_{li} = \frac{(2^{\frac{\mu s_i}{t_{d,i} B}} - 1)\Gamma_2 \sigma_{2,i}^2}{PN\gamma_i} \quad (10)$$

Replacing these expressions into (5) and (8), the total energy consumption by all users can be written as

$$E_u = \sum_{i=1}^{K} \left[\frac{t_{u,i}(2^{\frac{s_i}{\nu t_{u,i} B}} - 1)\Gamma_1 \sigma_{1,i}^2}{N\gamma_i} + \kappa_i c_i (u_i - s_i) f_{u,i}^2 \right] \quad (11)$$

Similarly, based on equations (7) and (9), the total energy consumption at the MEC server for computation offloading is

$$E_m = \sum_{i=1}^{K} \left[\frac{t_{d,i}(2^{\frac{\mu s_i}{t_{d,i} B}} - 1)\Gamma_2 \sigma_{2,i}^2}{N\gamma_i} + \kappa_m d_m f_{mi}^2 s_i \right] \quad (12)$$

The energy minimization problem for computation offloading can then be given as

$$(P_{\text{CO}}): \min_{\boldsymbol{s},\boldsymbol{t}} E_{\text{total}} = (1-w)E_u + wE_m \quad (13)$$

$$\text{s.t. Eqs. } (11) - (12) \quad \text{(a-b)}$$

$$\sum_{j=1}^{3} (T_j) \leq T_d, \quad \text{(c)}$$

$$\frac{c_i(u_i - s_i)}{f_{u,i}} + t_{u,i} - T_d \leq 0 \ \forall i \in [1, K] \quad \text{(d)}$$

$$t_{u,i} - T_1 \leq 0, \quad \forall i \in [1, K] \quad \text{(e)}$$

$$t_{d,i} - T_3 \leq 0, \quad \forall i \in [1, K] \quad \text{(f)}$$

$$\frac{d_m s_i}{f_{mi}} - T_2 \leq 0 \quad \forall i \in [1, K] \quad \text{(g)}$$

Here E_{total} is weighted sum of energy consumed at all users (E_u) and the MEC (E_m), with $1-w$ and w as the respective weights. The optimizing variables of this problems are time allocation $\boldsymbol{t} = [t_{u,1}...t_{u,K}, t_{d,1}...t_{d,K}, T_1, T_2, T_3, T_c]$, and offloaded data $\boldsymbol{s} = [s_1...s_K]$. Given parameters of the problems are T_d as the total latency constraint, P as the AP's transmit power, B as the channel bandwidth, Γ_1, Γ_2 as the uplink and downlink capacity gaps, (κ_i, c_i) and (κ_m, d_m) as the switched capacitance and CPU cycle information at the users and the MEC respectively.

Constraints (a-b) show the total energy consumption at the users and the MEC respectively, which includes the energy consumed for offloading/downloading and computation. Constraints (c-d) represent the constraint that both the time consumed for all three phases at the MEC, and the time consumed for offloading $\boldsymbol{t_u}$ and local computation at each user $\boldsymbol{t_L}$ should not exceed T_d. Constraints (e-g) show that the time consumed separately for offloading $\boldsymbol{t_u}$, computation of users' tasks at the MEC $\boldsymbol{t_M}$, and downloading time $\boldsymbol{t_d}$ for each user's results must be less than the maximum allowable time, $\{T_1, T_2, T_3\}$, for that phase as given in {(5),(9), (7)} respectively.

B. Maximization Of Received Energy By Wireless Charging

The above offloading problem is followed by the wireless charging problem given below

$$(P_{\text{WC}}): \max_{\boldsymbol{W_q}} \sum_{i=1}^{K} \xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c \quad (14)$$

$$\text{s.t. } \text{tr}(\boldsymbol{W_q}) \leq P \quad \text{(a)}$$

$$\xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c \leq e_i \ \ \forall i = 1...K \quad \text{(b)}$$

Here the charging time is defined as $T_c = T_d - T_1^\star - T_3^\star$, where $T_1^\star$ and $T_3^\star$ are the optimal time allocation for offloading and downloading operations obtained by solving (P_{CO}). In this way, the two problems are formulated in a sequential manner in compliance with the overall latency constraint. The charging time T_c denotes that wireless charging occupies all the time within T_d outside the data transmission operations of offloading and downloading. The optimizing variable is the beamforming matrix for wireless charging $\boldsymbol{W_q} \in \mathbb{R}^{N \times N}$. The objective function is a sum of the received energy for all users and the objective is to maximize this overall received energy at the users. Constraint (a) represents the physical layer constraint on the maximum transmission power of the AP. Constraint (b) shows that the amount of received (charged) energy at the i^{th} user is no more than the energy that it requests.

Problem (P_{CO}) is a semi-definite programming problem where the objective function and constraints are linear trace functions of $\boldsymbol{W_q}$ and hence convex. We can show that strong duality holds since Slater's condition is satisfied, that is, we can find a strictly feasible point ($\boldsymbol{W_q} = p\boldsymbol{I}_{N \times N}$, $p \leq P/N$) in the relative interior of the domain of the problem where the inequality constraints hold with strict inequalities [36].

IV. Data Partitioning And Time Allocation for Computation Offloading

A. Problem Analysis

In this section we analyze the computation offloading problem (P_{CO}) and show that it can be decomposed into simpler problems. The multivariable problem in (13) is a non-linear and non-convex optimization problem. Following a similar approach as in [5], the objective function f_0 for (P_{CO}) is a convex function of s_i. Furthermore, provided that the gradient of $f_0(\cdot)$ with respect to s_i evaluated at $s_i = 0$ is positive, which is often satisfied in typical network settings, then the

total energy in problem (P_{CO}) is an increasing function of each s_i and there exists an optimal point, $s_i^\star \ \forall i \in [1, K]$, which minimizes E_{total} within the latency constraint. If offloaded data $\boldsymbol{s}$ is fixed, then problem (P_{CO}) turns out to be convex in the remaining variables as stated in the following lemma. Lemma 1 lets us decompose the original non-convex problem (P_{CO}) into simpler convex subproblems which will be used in the subsequent algorithm design.

Lemma 1. *For a given set of offloaded data $\boldsymbol{s}$, the problem (P_{CO}) is convex in the time allocation variable $\boldsymbol{t}$.*

Proof. Proof follows by examining each constraint and showing that with fixed s_i, it is a convex function. Details in Appendix A. □

Since CPU frequencies are not optimizing variables, for given s_i in (P_{CO}), we can find in closed form the optimum time consumed by the MEC to compute each user's tasks, and the overall time T_2 spent for the data computation function at the MEC as in Lemma 2 next.

Lemma 2. *For a given value of the offloaded data s_i, the computation time for the offloaded data T_2 can be predetermined in closed form as follows*

$$T_2 = \max_i \frac{d_m s_i}{f_{mi}} \tag{15}$$

and hence constraint (13g) can be excluded from the problem (P_{CO}).

Proof. Directly from constraint (g) in (13) for a given s_i. □

B. Optimal Primal Solution

Next we present the solution for the optimal time allocation for the computation offloading problem (P_{CO}). Since the problem is convex based on Lemma 1, we adopt a primal-dual solution using the Lagrangian duality analysis similar to that proposed in [5] and derive the optimal solution as given in Theorem 1 below.

Theorem 1. *The offloading and downloading time, $t_{u,i}$ and $t_{d,i}$ respectively, can be obtained as a solution of the form*

$$x = \frac{cB}{\ln 2}\Big(W_0\Big(\frac{-y}{\sigma^2 e} - \frac{1}{e}\Big) + 1\Big) \tag{16}$$

where $y = -\frac{\beta_i+\theta_i}{(1-w)}$, $x = x_{1,i} = \frac{1}{t_{u,i}}$, $c = \frac{\nu}{s_i}$, $\sigma^2 = \frac{\Gamma_1 \sigma_{1,i}^2}{N\gamma_i}$ to solve for $t_{u,i}$, and $y = \frac{-\phi_i}{w}$, $x = x_{2,i} = \frac{1}{t_{d,i}}$, $c = 1/\mu s_i$, and $\sigma^2 = \frac{\Gamma_2 \sigma_{2,i}^2}{N\gamma_i}$ to solve for t_d, i. Here θ_i, β_i and ϕ_i are the dual variables associated with the constraints (d), (e) and (g) of problem (P_{CO}) in (13) respectively.

Proof. The solution in (16) can be obtained directly by applying KKT conditions on the Lagrangian dual of the problem P_{CO} with respect to $t_{u,i}$ and $t_{d,i}$. Detailed proof can be obtained using an approach similar to that in [5, Theorem 1] and is omitted for brevity. □

V. Energy Beamforming for Wireless Charging

A. Sequential And Nested Algorithm Structures

In this section, we derive the solution for the optimal transmit covariance matrix, $\boldsymbol{W_q}$ by finding the optimal energy beam directions and also the optimal beam power allocation. For the received energy maximization problem (P_{WC}), we use Lagrangian duality analysis to obtain the optimal beam directions as described in Theorem 2 below.

Theorem 2. *For maximizing the received energy, the optimal directions for energy beams are $\boldsymbol{U_q^\star} = \boldsymbol{U_C}$, where $\boldsymbol{U_C}$ is obtained from the eigenvalue decomposition of $\boldsymbol{C} = \boldsymbol{U_C \Lambda_C U_C^*}$, such that $\lambda_{C,1} \geq \lambda_{C,2} \geq \ldots \geq \lambda_{C,N}$, where*

$$\boldsymbol{C} = \chi \boldsymbol{I} + T_c \sum_{i=1}^{K} \xi_i (1+\rho_i) \boldsymbol{h_i h_i^*} \tag{17}$$

Here χ and ρ_i are the dual variables associated with constraint (14a) and the i^{th} constraint in (14b) respectively.

Proof. See Appendix B. □

Theorem 2 provides the optimal directions of the energy beams for the beamforming matrix, $\boldsymbol{W_q}$. What is left now is to obtain the optimal power allocation across the energy beams, that is, the eigenvalues of the transmit covariance matrix for wireless charging. To this end, we substitute the optimal beam directions from Theorem 2 into (P_{WC}) and rewrite the formulation in terms of the beam power allocation only as (P_{BP}) below. Beam power allocation, $\boldsymbol{\lambda_q}$, can then be obtained as a solution to a Linear Programming (LP) problem given in Theorem 3 below.

Theorem 3. *The optimal beam power allocation which maximizes the received energy through wireless charging is derived as a solution of the LP problem below*

$$(P_{BP}): \max_{\boldsymbol{\lambda_q}} \ \sum_{i=1}^{K} \boldsymbol{d_i^* \lambda_q} \tag{18}$$

$$s.t. \ \sum_{i=1}^{K} \lambda_{q,i} \leq P, \qquad \lambda_{q,1} \geq \ldots \geq \lambda_{q,K} \geq 0 \ \text{(a-b)}$$

$$\boldsymbol{D \lambda_q} \leq \boldsymbol{b} \ \text{(c)}$$

where $\boldsymbol{\lambda_q} = [\lambda_{q,1}, \ldots, \lambda_{q,K}]^T$, $\boldsymbol{D} \in \mathbb{R}^{K \times K} = [\boldsymbol{d_1^*} \ldots \boldsymbol{d_K^*}]$, $\boldsymbol{d_i}^* = \text{diag}(\boldsymbol{r_i r_i^*})$, $\boldsymbol{r_i^*} = \boldsymbol{h_i^* U_C} = \boldsymbol{h_i^* U_q}^\star$ and $\boldsymbol{b} \in \mathbb{R}^{K \times 1} = [\pi_1 \ldots \pi_K]$, $\pi_i = \frac{e_i}{\xi_i T_c} \ \forall i = 1 \ldots K$.

Proof. Obtained by substituting optimal beam directions from Theorem 2 in (P_{WC}). Details in Appendix C. □

Note that in the above solutions for (P_{WC}), since the goal is energy maximization, the eigenvalues of $\boldsymbol{W_q}$ and $\boldsymbol{C}$ are of the same order. All the eigenvectors of each matrix are ordered according to their corresponding eigenvalues. The optimal solutions derived thus far are specific to the respective problems (P_{CO}) and (P_{WC}), and thus reveal the optimal solution structure that otherwise would be obscured by using a generic solver. Next we use these optimal solutions to design customized algorithms to solve these problems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

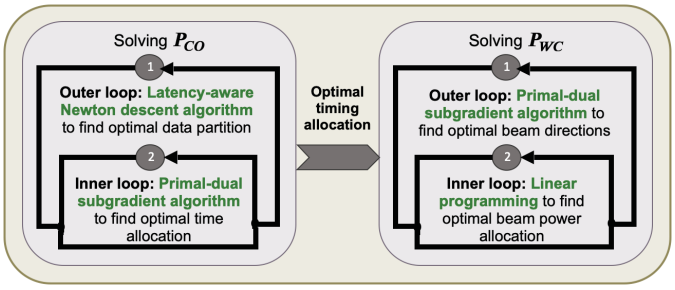

Figure 3: Structure of the main algorithm Alg. 1, consisting of sequential sub-algorithms for solving P_{CO} and P_{WC}.

VI. Algorithm Design

In this section, we discuss the algorithm structure to solve the two sequentially formulated problems P_{CO} and P_{WC}. Based on the way these two problems are formulated, P_{CO} will be solved first to obtain the optimal data partitioning and time allocation for computation offloading. This optimal time allocation will then be used in P_{WC} as a given parameter in order to find the optimal energy beamforming structure.

The algorithm for solving (P_{CO}) is designed based on Lemma 1 to have a nested architecture with an outer and an inner loop, in which the outer loop solves for s_i decrementally while the inner loop solves for $\boldsymbol{t}$ at a fixed value of s_i. Specifically, the nested algorithm works as follows. We first initialize the offloaded bits $\boldsymbol{s}$ and the dual variables in the outer algorithm. At the current value of s, the inner algorithm is executed, for which we use a primal-dual approach employing a subgradient method. At convergence where the stopping criterion for the dual problem is satisfied, the inner algorithm returns the control to the outer algorithm. Based on the newly updated primal solution from the inner algorithm, we proceed to updating $\boldsymbol{s}$ by some Δs_i for each user for the next outer-loop iteration, using a latency aware descent algorithm. Similar to [5], the latency aware descent algorithm is based on the standard Newton method, with a novel modification to the classical stopping criterion to account for the latency constraint.

Problem (P_{WC}) solves for the transmit covariance matrix $\boldsymbol{W_q}$ as an independent problem after obtaining the optimal time allocation solution from (P_{CO}) to calculate the charging time T_c as $T_c = T_d - T_1^\star - T_3^\star$. The algorithm for solving (P_{WC}) also has a nested structure, with an outer algorithm to establish the optimal beam directions and an inner algorithm for the beam power allocation. Specifically, at each iteration of (P_{WC}), an outer algorithm step finds the optimal dual variables for the beam direction solutions in Theorem 2 via a subgradient method, and calls to an inner algorithm which solves the LP problem (P_{BP}) in Theorem 3 for the optimal beam power allocation using a standard convex solver. Once the beam power allocation is found, the inner algorithm returns to the outer one in order to update the dual variables, and the process continues until convergence is reached in the outer algorithm. In the case of (P_{WC}), the outer algorithm is primal-dual, and the inner algorithm is linear programming. The algorithm flow is depicted in Figure 3 and steps for solving both problems (P_{CO}) and (P_{WC}) are given in Algorithm 1.

Algorithm 1 Solution for (P_{CO}) and (P_{WC})

Given: Distances d_i $\forall i$. Channel $\boldsymbol{H} = \boldsymbol{G}^{\boldsymbol{T}}$. Precision, ϵ_1, ϵ_2, Data u_i, Latency T_d. Initialize: s_i

Begin Sub-algorithm for (P_{CO})

Outer Loop (Latency-aware Newton Method): **Repeat**

1) Compute $\Delta \boldsymbol{s}$ using the Newton method, where
$$\Delta \boldsymbol{s} := -\nabla^2 f_0(\boldsymbol{s})^{-1} \nabla f_0(\boldsymbol{s})$$
and $f_0(.)$ is the objective function in (13)
2) *Inner Loop (Subgradient Method)*
   - Calculate $t_{u,i}$ and $t_{d,i}$, using (16). Then $T_1^\star = \max t_{u,i}^\star$ and $T_3^\star = \max t_{d,i}^\star$.
   - Update p_i and η_i using (10) and calculate $\sigma_{1,i}^2$ and $\sigma_{2,i}^2$.
   - Solve the dual problem in (20):
     - Establish the dual function in (19) by using Theorem 1
     - **Repeat**
       a) Compute subgradients in (22a-d)
       b) Update dual-variables using subgradient method
       Until dual subgradients converge with ϵ_2 as in (23)
3) *Line search and Update*. $s_i := s_i + t_i \Delta s_i$.

Until stopping criterion for Newton method is satisfied: $\lambda^2/2 < \epsilon_1$ or latency constraint T_d is met. $\lambda := -\nabla f_0(\boldsymbol{s})^T \Delta \boldsymbol{s}$

End Sub-algorithm for (P_{CO})

Given: Optimal time allocation from (P2) in Step 2 above, find $T_c = T_d - T_1^\star - T_3^\star$

Begin Sub-algorithm for (P_{WC})

Outer Loop (Subgradient Method): **Repeat**

1) Solve for beam directions $\boldsymbol{U_C}$ as function of dual variables as in Theorem 2
2) *Inner Loop (Linear Programming):*
   - Solve for beam power allocation $\boldsymbol{\lambda_q}^\star$ as an LP (P_{BP}) in Theorem 3
3) Update primal variable
$$\boldsymbol{W_q}^\star = \boldsymbol{U_C} \boldsymbol{\Lambda_q^\star} \boldsymbol{U_C^*}, \text{ where } \boldsymbol{\Lambda_q^\star} = \mathbf{diag}(\boldsymbol{\lambda_q}^\star)$$
4) Update the dual problem in (21):
   - Establish the dual function in (26) by using Theorems 2 and 3
   - Compute subgradients in (22e-f)
   - Update dual-variables using subgradient method

Until dual subgradients converge with ϵ_2 as in (23)

End Sub-algorithm for (P_{WC})

A. Primal-Dual Algorithms

For the inner optimization in (P_{CO}) and the outer algorithm in (P_{WC}), we design primal-dual algorithms where the primal variable are obtained as closed-form functions of the dual variables, which are found by solving the dual problem using a sub-gradient method. The dual-function for the convex

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

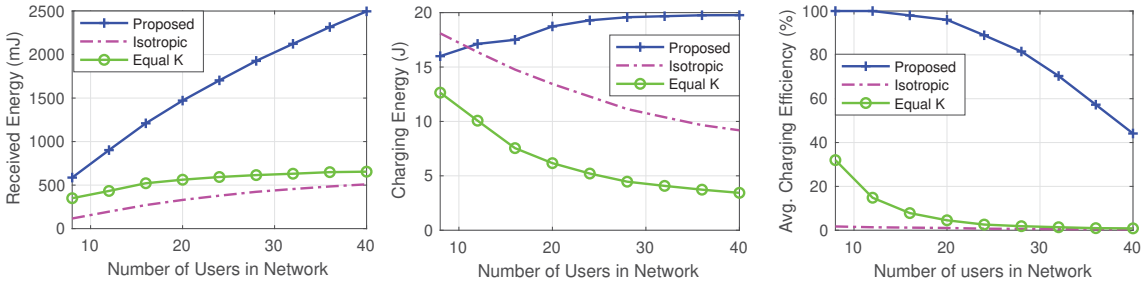

Figure 4: Comparison of the proposed wireless charging scheme with isotropic wireless charging, and directed K-beam charging with equal power allocation (equal K): (left) amount of total received (charged) energy; (middle) total transmitted energy consumption; (right) the average charging efficiency defined as in (24).

optimization problem (P_{CO}) at a given s_i can be defined as

$$g_{\text{CO}}(\lambda_1, \boldsymbol{\beta}, \boldsymbol{\xi_i}, \boldsymbol{\phi}) = \inf_{\boldsymbol{t}} \mathcal{L}_{\text{CO}}(\boldsymbol{t}, \lambda_1, \boldsymbol{\beta}, \boldsymbol{\xi_i}, \boldsymbol{\phi}) \tag{19}$$

where $\mathcal{L}_{CO}$ is the Lagrangian for problem (P_{CO}). The optimal value of $\boldsymbol{t}$ which minimizes this Lagrangian is given in Theorem 1, based on which the dual function can be obtained. Then the dual-problem for (P_{CO}) is defined as

$$\begin{aligned} P_{\text{CO}}\text{-dual:} \max \;\; & g_{\text{CO}}(\lambda_1, \boldsymbol{\beta}, \boldsymbol{\xi_i}, \boldsymbol{\phi}) \\ \text{s.t. } & \lambda_1 \geq 0, \beta_i, \theta_i, \phi_i \geq 0 \;\forall i = 1...K \end{aligned} \tag{20}$$

where λ_1, $\boldsymbol{\beta}, \boldsymbol{\xi_i}$, and $\boldsymbol{\phi}$ are the dual variables associated with constraints (c-f) in (13), respectively.

Similarly by maximizing the Lagrangian in (25) for problem (P_{WC}), using the optimal $\boldsymbol{W_c^\star}$ as derived from Theorems 2 and 3, the dual function $g_{\text{WC}}(\boldsymbol{\rho}, \chi)$ for (P_{WC}) can be established as in (26), and its dual problem is given as

$$\begin{aligned} \text{P}_{\text{WC}}\text{-dual:} : \min \;\; & g_{\text{WC}}(\boldsymbol{\rho}, \chi) \\ \text{s.t. } & \chi \geq 0, \rho_i \geq 0 \text{ for } i = 1...K \end{aligned} \tag{21}$$

Using the closed form expressions for the optimal primal variables in terms of the dual-variables in Theorems 1-3, the dual functions above are functions of only the dual-variables.

The subgradient terms with respect to all dual variables of original problems (P_{CO}) and (P_{WC}) are as given below

$$\nabla_{\lambda_1} \mathcal{L} = \sum_{j=1}^{3} T_j - T_{\text{delay}} \tag{22a}$$

$$\nabla_{\beta_i} \mathcal{L} = t_{u,i} - T_1, \tag{22b}$$

$$\nabla_{\phi_i} \mathcal{L} = t_{d,i} - T_3, \tag{22c}$$

$$\nabla_{\theta_i} \mathcal{L} = \frac{c_i q_i}{f_{u,i}} + t_{u,i} - T_d, \tag{22d}$$

$$\nabla_{\rho_i} \mathcal{L} = \xi_i \text{tr}\left(\boldsymbol{h_i^* W_q h_i}\right) T_c - e_i, \tag{22e}$$

$$\nabla_{\chi} \mathcal{L} = \text{tr}(\boldsymbol{W_q}) - P \tag{22f}$$

For implementation of the primal-dual algorithms, we use the subgradient method to solve the constrained convex optimization problems (P_{CO}) and (P_{WC}) [37]. The designed algorithms find the subgradients for the negative dual function $-g_{\text{CO}}$, since the dual problem in (20) is a maximization problem for the dual function, and for the positive dual function g_{WC}, since the dual problem in (21) is a minimization problem. At each iteration, the primal variables are updated based on Theorems 1-3. The dual variables vector x is updated as $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \beta_k \boldsymbol{g}^{(k)}$, where β_k is the k^{th} step-size, and $\boldsymbol{g}^{(k)}$ is the subgradient vector at the k^{th} iteration evaluated using the sub-gradient expressions in (22a-f). We use the non-summable diminishing step size, setting $\beta_k = 1/\sqrt{k}$. Since the subgradient method is not a descent method, the algorithms keep track of the best point for the dual functions at each iteration of the inner algorithm. These primal-dual update steps are repeated until the desired level of precision, ϵ_2, is reached for the stopping criterion. In the subgradient method, since the key quantity is not the function value but rather the Euclidean distance to the optimal set [37], therefore, for our implementation we define the stopping criterion as

$$\|\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}\|_2 \leq \epsilon_2. \tag{23}$$

B. Algorithm Complexity

For (P_{CO}), the outer algorithm is a latency-aware descent algorithm based on the Newton method and solves the optimization problem $\min_{\boldsymbol{s}} f(\boldsymbol{s})$, where $f = E_{\text{total}}$, and $\text{dom}(f) \in \mathbb{R}^K$. For K users and $\boldsymbol{s} \in \mathbb{R}^{K \times 1}$, the computation cost for each Newton iteration requires $\mathcal{O}(K^3)$ flops [38] and the backtracking line search requires $\mathcal{O}(K)$ flops per inner backtracking step. The novel latency-aware stopping criterion is a max operation over K users, with complexity $\mathcal{O}(K)$ [5]. The inner algorithm is based on the subgradient method where we use the non-summable diminishing step size for which the algorithm is guaranteed to converge to the optimal value with a theoretical iteration complexity of $\mathcal{O}(1/\epsilon^2)$ [37] [39]. The chosen stopping criterion of the inner subgradient algorithm is a norm calculation which requires $2(3K+1)$ flops based on the size of the subgradient vector $\boldsymbol{g}^{(k)}$ for (P_{CO}).

For (P_{WC}), in the outer algorithm, the most computationally intensive step is finding the optimal beam directions for an N-antenna massive MIMO array which requires SVD of $\boldsymbol{C} \in \mathbb{C}^{N \times N}$ with a computation cost of $\mathcal{O}(N^2)$. The inner algorithm is to solve a linear programming problem (P_{BP}), where an LP has complexity class P. Finally $\boldsymbol{W}$ is obtained through matrix multiplication in which we take into account the zero-elements of λ_q and hence it has complexity $\mathcal{O}(N^2K^2)$. The chosen stopping criterion for the outer algorithm is again a norm calculation which requires requires $2(K+1)$ flops based on the size of the subgradient vector $\boldsymbol{g}^{(k)}$ for (P_{WC}).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

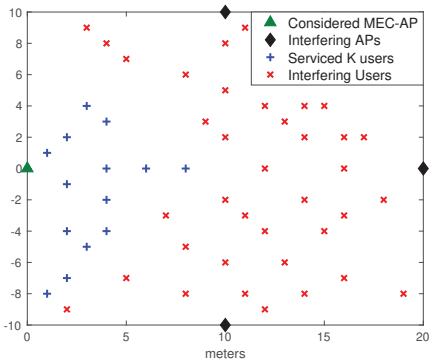

Figure 5: Simulation network layout of a conference exhibition hall with 4 AP-MEC serving multiple and randomly located users. The number of active users requesting computation offloading and wireless charging varies between 16-40.

VII. Numerical Results

In this section, we evaluate the solution of the sequential problem with respect to energy and time consumption, the partition of bits offloaded to the MEC for computation and the received energy via wireless charging. We consider a $20\text{m} \times 20\text{m}$ area (typical service area for AR applications with bi-directional transmission [3]) with 4 MEC-APs, each with $N = 100$ antennas as shown in Figure 5. We start with 16 users randomly located in the network with $K = 4$ users per AP's coverage area, and increase the number of active users up to 40 users ($K = 10$ per AP area) in various simulations. For simulations, $w = 10^{-3}$, $T_d = 20$ms (for AR/VR applications [40]), $B = 5$MHz, $\tau_c = BT_d$, $\Gamma_1 = \Gamma_2 = 1.25$, $\mu = 2$, $\kappa_i = 0.5$pF, $\kappa_m = 5$pF, $c_i = 1000$, $d_m = 500$, $\gamma = 2.2$, $\sigma = 2.7$dB, $\sigma_r^2 = -127$dBm, $\sigma_k^2 = -122$dBm, $f_{u,i} = f_u = 1800$ MHz $\forall i$. Each MEC processor has 24 cores with maximum frequency of 3.4GHz, and we use $f_{m,i} = f_m = \frac{24\times 3400}{K}$ MHz $\forall i$. Transmit power available at user and AP is 23 dBm and 46 dBm respectively. To calculate the interference and noise power ($\sigma_{1,i}^2$, $\sigma_{2,i}^2$) which include massive MIMO pilot contamination and intercell interference, we assume that user terminals transmit at their maximum power, that is $p_{qi} = 23$dBm, and the interfering APs use equal power allocation in the downlink, that is $\eta_{qi} = \frac{1}{K}$ $\forall i$. Numerical results are averaged over 100 independent channel realizations of $\mathbf{H}$ and $\mathbf{G}$. The results in Figures 4, 7, 12, 13 and 14 are averaged over 200 spatial realizations (randomly generated user locations).

A. Comparison of Wireless Charging Schemes

Figure 4 shows a comparison of the proposed maximization wireless charging scheme with two other schemes: (i) isotropic scheme where $\boldsymbol{W_q} = \frac{P}{N}\boldsymbol{I}$ and equal charging power P/N is allocated across all N antennas of the AP, and (ii) equal K with directional charging using the beamforming directions proposed in Theorem 2, but with equal power allocation P/K across K energy beams. For fairness of comparison with the sequential scheme, we use power scaling for the other two schemes such that each user only receives an amount of energy at most equal to requested, similar to the sequential scheme. Since wireless charging is proposed as a billable service for future networks, this is also a necessary design consideration from the service providers' and consumers' perspectives.

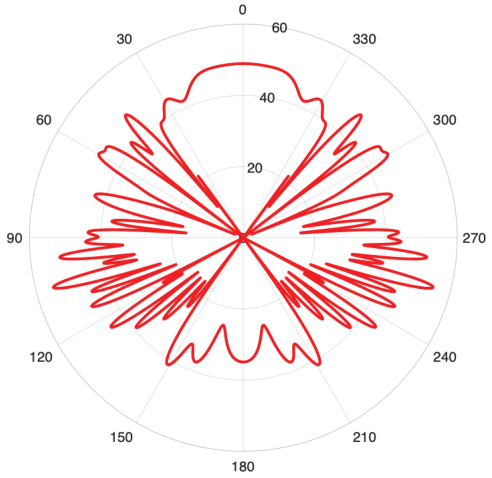

Figure 6: A typical wireless charging beam pattern for simultaneously charging multiple UEs from an MEC-AP, where shown is the strongest beam out of 10 beams for this channel realization

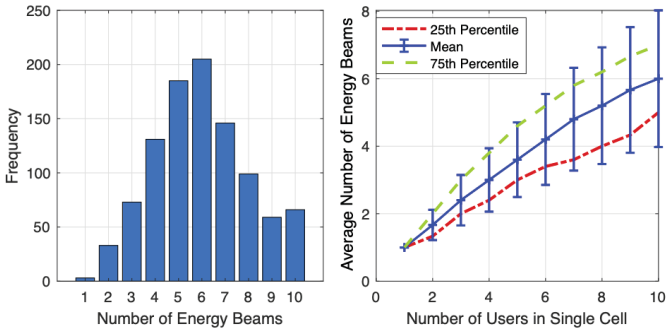

Figure 7: Distribution of the number of charging beams for K = 10 users per cell (left), and the average number of charging beams vs the number of users in each cell (right).

Figure 4 shows the received energy on the left, the transmitted energy in the middle, and the average charging efficiency on the right. Average charging efficiency (per time block) is defined as the average percentage of received energy, in the q^{th} time block as denoted by (q), at the users end compared to the requested energy, given as

$$\text{Avg. Charging Efficiency } (\%) = \frac{\sum_{i=1}^{K} \xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c^{(q)}}{\sum_{i=1}^{K} e_i^{(q)}} \tag{24}$$

Note that the requested energy at the q^{th} time block excludes the amount of energy requests already fulfilled in the previous time block(s). As illustrated in this figure, the sum received energy for the energy maximization sequential scheme is significantly larger than the other two schemes. Beamforming with equal power allocation scheme performs better than the isotropic scheme, since it consumes lesser charging energy and still delivers higher energy to the users. Comparing the average charging efficiency for all the schemes, however, the wireless charging maximization scheme enables substantially higher charging efficiency. The average efficiency is seen to decrease with an increase in the network size as expected.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

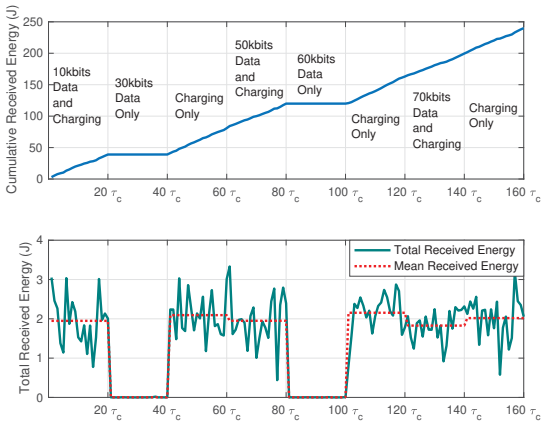

Figure 8: A typical example of system performance for K users over multiple time-slots under three modes (a) Computation and Charging, (b) Computation only, and (c) Charging only

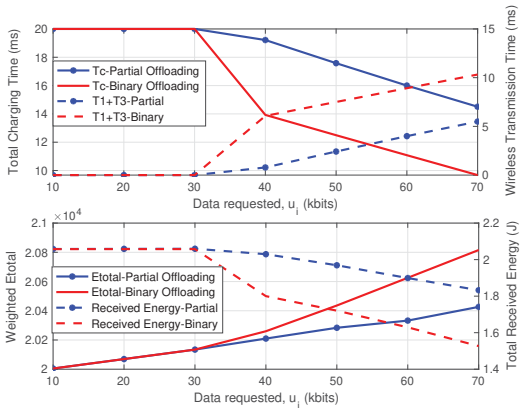

Figure 9: Comparison of time and energy consumption between proposed partial and binary offloading schemes, showing more energy and time efficiency when data can be split for partial offloading.

B. Charging Beams

Figure 6 shows a typical beam radiation pattern of the proposed scheme for the simultaneous wireless charging of all active UEs requesting charging. We see that in addition to the main lobe, there are a large number of side lobes where the nulls are not as deep, which allows for increased charging energy levels to users. This "*null-fill*" property is a common design feature to alter the energy distribution for the various antenna elements in the array [41]. For maximal received energy, users may receive wireless charging not only from the main beam but also from the side lobes and backscattering which can be an important consideration for wireless charging.

Figure 7 shows the distribution of the optimal number of energy beams for $K = 10$ users per cell (left) and the average number of beams for an increasing number of users in the network (right). In comparison, for the isotropic wireless charging, there are always $N > K$ energy beams. For the case of K beams with equal power allocation, the number of beams is equal to the number of users in the cell. While multiple energy beams may be necessary for a multi-user system as also previously discussed in [13], the optimal number of energy beams for the proposed wireless charging scheme is usually less than the number of users. Since each energy beam can charge multiple users simultaneously, the transmit beamforming can be intelligently designed as proposed to limit the number of energy beams which can prevent energy losses caused by transmitting energy in numerous directions. Therefore, for the proposed received energy maximization, the optimal number of beams on the average is much lower than the number of users.

C. Charging Profile

Figure 8 shows a typical example of the system's charging performance over time under the three modes of operation, namely, data and charging, data only, and charging only. The charging only and data only modes are special cases/subsets of the data and charging mode. For the joint data and charging mode, both data and energy requests are non-zero, that is $u_i > 0, e_i > 0 \ \forall \ i$. For the data only mode, $e_i = 0$ and for the charging only mode, $u_i = 0$. The figure shows the time profile for the received energy by the users. The time axis is plotted in terms of the coherence interval τ_c, to show that the energy values are calculated for a new channel realization after every coherence interval which corresponds to the variation in the received energy value over time in the bottom plot. For the results shown we assume $\tau_c = BT_d$, with $T_d = 20$ms.

The cumulative energy on the top figure show that during the data only operation in which no users request wireless charging, there is no increase in the charged energy as expected. Correlating with the bottom plot, we see a decrease in the mean received energy during the joint phase of computation and power transfer (data and charging) as compared to the charging only phase. The results verify that our algorithm works as expected since with computation, a portion of time from τ_c is spent on data computation and wireless transmission, as compared to the charging only mode where the entire duration is spent for wireless charging. The cumulative top plot show that over an extended period of time over both computation and non-computation intervals, wireless charging can deliver a significant amount of energy.

D. Effect of the Amount of Data Requested and Partitioning

Figure 9 shows a comparison of the proposed partial offloading scheme, where data partitioning is used to divide the computation between the MEC and each user, with the binary offloading scheme where the task is atomic and is either offloaded or computed locally as a whole. We compare the time and energy consumption for the two schemes as the amount of data requested is increased under a fixed latency constraint of $T_d = 20$ms. To evaluate the solution for the binary offloading scheme, we consider all possible binary offloading combinations and choose the one with the lowest overall energy consumption.

We see significant disparity between the binary and partial offloading schemes when large amounts of data are requested. For low data requests, local processing at users is optimal so both schemes consume the same energy and the entire duration is spent for wireless charging by the MEC concurrently with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

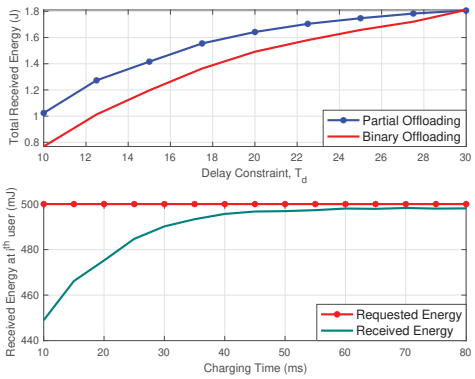

Figure 10: Amount of charged (received) energy comparison between binary offloading and the proposed partial offloading scheme: (top) for all K=4 users, each requesting 500mJ of energy, (bottom) for an individual user requesting 500mJ

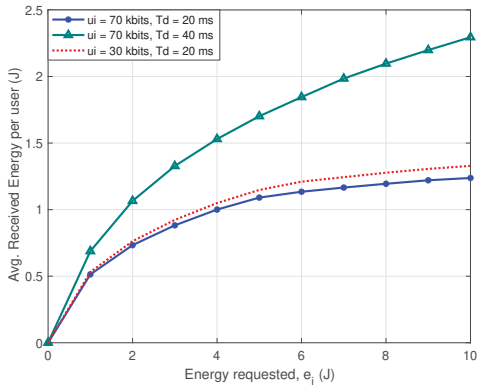

Figure 11: Amount of received energy by a user as it requests more energy under different computation and time requirements

the local computation at the users. For larger data requests, however, the binary offloading scheme spends far less time for charging, since the time for wireless transmission to offload all the data to the MEC is greater. Owing to this increased time for wireless transmission in the binary scheme, the overall weighted system energy consumption is much larger than that of partial offloading. Partial offloading not only results in a lower overall weighted energy consumption, but also leads to higher received energy at the users during wireless charging by the MEC because of the longer charging time. Partial offloading with data partitioning therefore appears as a potent design variable for the resource allocation problem, with significant impact on the wireless charging capability of the system.

E. Effect of the Latency Constraint and Charging Time

Figure 10 (top) shows the total received or charged at the users, each requesting 500mJ of energy, as the delay constraint is relaxed, that is, T_d is increased at a fixed amount of requested data, $u_i = 50$ kbits. For this amount of data, binary offloading results in lower received energy since all data is offloaded to the MEC to meet the latency requirement. This results in larger time consumption for wireless transmission, consequently reducing the charging time and hence the charged energy. For relaxed latency, however, both binary and partial offloading schemes compute data locally, and hence the plots converge.

Figure 10 (bottom) shows the received energy, that is the amount of charge the user receives through wireless power transfer, as the charging time is increased. We show the requested and received energy for one user in a 16 user network, where each user requests 500 mJ of energy from the MEC, and the network is in *charging only* mode, that is, the users do not request any data for computation. For longer charging times, the MEC fulfills the user's demand for wireless charging almost completely.

F. Effect of the amount of Energy Requested

Figure 11 shows the amount of energy received by a user through wireless charging, as the amount of energy requested by the user is increased in the *data and charging* mode, that is the users jointly request data computation and wireless charging. We assume all the users requesting the same amount of energy, that is $e_i = e \ \forall i$. For lower amounts of requested energy, we see that the MEC-AP strives to fulfill the energy demand to a large extent, however, as the energy demands are increased by all the users simultaneously, the wireless charging by the MEC-AP cannot cope with the wireless charging demand in full.

We compare the amount of energy received through wireless charging for three scenarios (a) all users request 70 kbits of data for computation, that is $u_i = 70$ kbits $\forall i$ under a latency requirement of $T_d = 20$ms, (b) $u_i = 70$ kbits with relaxed latency $T_d = 40$ ms, and (c) reduced data request $u_i = 30$ kbits at $T_d = 20$ ms. Even with less than half of the amount of data request, the amount of wireless charged energy increases only slightly, showing that the amount of data for offloading (while feasible) has a small impact wireless charging. On the other hand, a twice-relaxed latency constraint has a significant impact on wireless charging by increasing the charged energy substantially.

G. Effect of Network Size

Figure 12 shows the total amount of energy received by all users during the wireless charging function, as the number of users in the network is increased, each requesting $e_i = 500$mJ

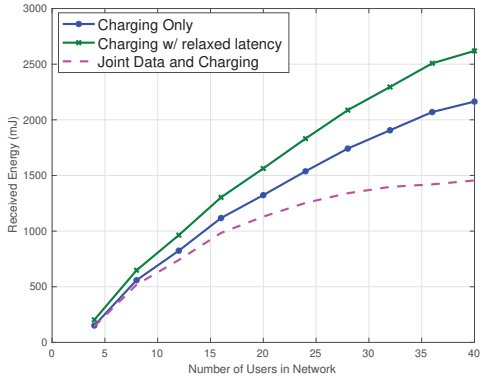

Figure 12: Amount of charged energy received at all users, each requesting 500mJ of energy, with and without data computation as the number of users in the network increases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

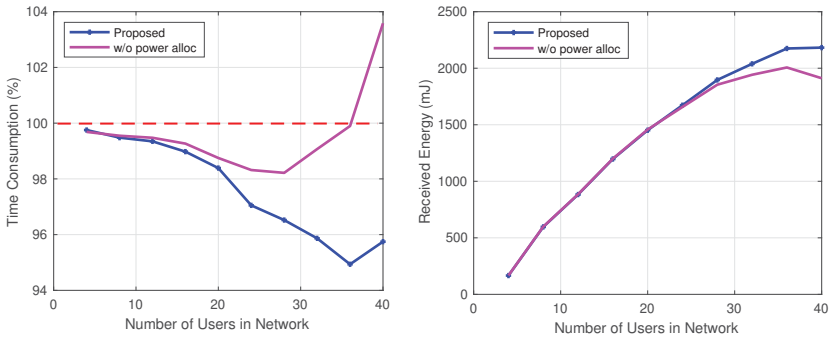

Figure 13: Effects of MEC transmit power control for computation offloading on (left) the time consumption of computation offloading, showing the scheme without power control starts violating the latency as the number of users increases; and (right) the amount of energy received via wireless charging.

of energy. We compare the charged energy under different scenarios, namely (i) *Charging Only* where each user only has energy requests and no data to offload, (ii) *Charging w/ relaxed latency* with charging-only mode but the latency constraint is relaxed from 20 ms to $T_d = 40$ ms, and (iii) *Joint Data and Charging* where each user request $u_i = 70$ kbits of data for computation along with its energy request. Either relaxing the latency constraint or having no data significantly increases the received energy during the wireless charging phase. The total received energy increases with the network size but exhibits a diminishing effect because of the total transmit power constraint.

H. Effect of MEC Transmission Power Allocation in Computation Offloading

Figure 13 also shows the effect of transmission power allocation in data transferring phases of computation offloading on the received energy as the network is increased, where each user requests $e_i = 500$mJ of energy and $u_i = 40$ kbits of data for computation. In our proposed scheme, transmit power control is implemented indirectly through the time and data partitioning, given by (10) via the optimized time allocation variables $\boldsymbol{t_u}$ and $\boldsymbol{t_d}$. In the scheme without power allocation, we fix the transmit power such that the MEC-AP allocates equal power for transmission beamforming to all users in downlink, and all users use the maximum transmit power available. Note that this is the power allocation for data transmission in offloading and not to be mistaken with power allocation for the energy beamforming discussed in Sec. VI.

An important finding for large network sizes is that without transmission power allocation, the network cannot cope with the data and energy requests within the latency constraint, evident by the percentage time consumption exceeding 100% for 35 or more users in the network. As the network size increases, transmit power allocation for computation offloading also has a positive impact on the amount of received energy via wireless charging. Transmit power control is only consequential for large network sizes and can be excluded from the optimization problem to reduce complexity in small networks.

I. Algorithm Convergence

Figure 14 shows, on the left, the convergence of the two algorithms solving optimization sub-problems P_{CO} and P_{WC} with $u_i = u = 10$kbits, $e_i = e = 1$J $\forall i$. The algorithm for P_{WC}, or P_{WC} *sub-algorithm* in short, based on nested subgradient method and linear programming converges in significantly fewer iterations compared to the algorithm for P_{CO}, or P_{CO} *sub-algorithm*, based on nested latency-aware Newton descent and subgradient methods. Not only does P_{WC} *sub-algorithm* converge in fewer iterations compared to the P_{CO} *sub-algorithm*, the time taken per iteration is also shorter for P_{WC} as shown in Figure 14 on the right.

The computation offloading P_{CO} *sub-algorithm* optimizes for data partitioning and time allocation for each user, leading to the number of optimizing variables for K users as $2K$. However, a key contributing factor to the increased number of iterations for P_{CO} is the size of the subgradient vector, which in this case is $\boldsymbol{g}^{(k)} \in \mathbb{R}^{3K+1}$ whereas for P_{WC}, $\boldsymbol{g}^{(k)} \in \mathbb{R}^{K+1}$. Since the subgradient algorithm is the most time consuming step in both sub-algorithms, the mean time per iteration for P_{CO} is larger than that for P_{WC} as seen in Figure 14.

Moreover, the wireless charging P_{WC} *sub-algorithm* calculates the beam directions for all K users through a single matrix factorization per iteration as in Theorem 2. The power allocation per energy beam is then solved via an efficient inner linear programming algorithm which scales slowly with the number of users in the network. For our implementation on a personal computer, the time step unit in Figure 14 is a second, however for faster machines, such as MEC servers, with the high-performance CPUs, this time-step may be significantly smaller.

VIII. Conclusion

We examined a massive MIMO enabled multi-access edge computing network providing computation offloading and on-request wireless charging to its connected users under a round trip latency constraint. We formulated a novel system-level problem to minimize the energy consumption for data offloading and to maximize the received energy from wireless charging, and design efficient algorithms to solve for data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3075920, IEEE Transactions on Wireless Communications

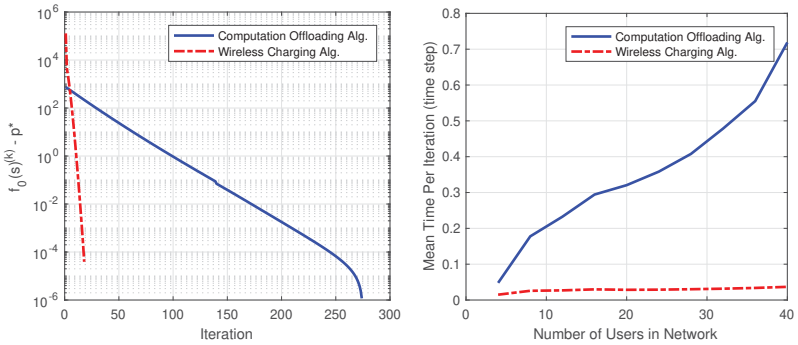

Figure 14: Convergence of the proposed sub-algorithms for P_{CO} and P_{WC}: (left) optimal objective accuracy vs number of iteration; (right) average execution time per iteration.

$$\begin{aligned}\mathcal{L}_{\text{WC}} &= T_c \sum_{i=1}^{K} \xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c + \chi \left(\text{tr}(\boldsymbol{W_q}) - P\right) + T_c \text{tr}\left(\left(\sum_{i=1}^{K} \xi_i \rho_i h_i h_i^*\right) \boldsymbol{W_q}\right) - \sum_{i=1}^{K} \rho_i e_i \\ &= T_c \text{tr}\left(\left[\chi \boldsymbol{I} + T_c \sum_{i=1}^{K} \xi_i (1+\rho_i) \boldsymbol{h_i h_i^*}\right] \boldsymbol{W_q}\right) - \sum_{i=1}^{K} \rho_i e_i - \chi P \\ &= T_c \text{tr}\left(\boldsymbol{C W_q}\right) - \sum_{i=1}^{K} \rho_i e_i - \chi P \end{aligned} \tag{25}$$

partitioning, time allocation and transmit energy beamforming matrices. Our algorithms demonstrated that data partitioning is a potent optimizing variable, as partial data offloading when possible leads to significant reduction in system energy consumption, lower transmission times and consequentially higher amount of received charged energy at users. On the other hand, MEC-AP transmit power allocation for downlink data transmission has little effect on the system energy consumption for small network sizes.

Our algorithm also illustrated that even with significant amounts of data to be computed, the network can deliver decent amounts of charged energy to the users over an extended period of multiple computation time-slots, therefore validating a practical coexistence of computation offloading and wireless charging. A comparison with isotropic power transfer and equal power energy beamforming shows that optimal design of the energy beamforming directions and beam power allocation in wireless charging is crucial for energy efficiency, and is necessary for adopting on-request wireless charging as a billable service for future networks.

IX. Appendix

A. Appendix A - Proof for Lemma 1

Consider problem (P_{CO}) in (13) at fixed values of s_i. The objective function is affine in E_u and E_m and hence convex.To show that the P_{CO} is convex in $\boldsymbol{t}$, we need to consider each constraint as follows.

- Constraints (c), (e), (f), (h) for (P) in (13) are linear in $\boldsymbol{t}$.
- For constraints (a) and (b), the first terms are of the form $f(x) = x2^{\frac{1}{x}}$ in $t_{u,i}$ and $t_{d,i}$ respectively, with $\nabla_x^2 f(x) = \frac{2^{\frac{1}{x}}}{x^3} > 0$ for $x > 0$, and hence $f(x)$ is convex in x.
- Relevant constraints are also linear and convex in E_u, E_m and T_j $\forall j$.

Based on the above, the objective is convex and all the constraints are convex in the remaining variables. Thus the problem is convex at given s_i. □

B. Appendix B - Proof for Theorem 2

To establish the optimal beamforming directions in terms of the dual variables, we analyze the Lagrangian function of problem (P_{WC}). Specifically, the Lagrangian for problem (P_{WC}) can be obtained as in (25), where χ and ρ_i are the dual variables associated with constraint (14a) and the i^{th} constraint in (14b), and the matrix $\boldsymbol{C}$ is defined as in (17). The dual-function for the problem (P_{WC}) can then be defined as

$$g_{\text{WC}}(\boldsymbol{\rho}, \chi) = \max_{\boldsymbol{W_q}} \mathcal{L}_{\text{WC}}(\boldsymbol{W_q}, \boldsymbol{\rho}, \chi) \tag{26}$$

We wish to find the beamforming matrix $\boldsymbol{W_q}$ to maximize the Lagrangian $\mathcal{L}_{\text{WC}}$. Applying the inequality relating trace of matrix product to the sum of eigenvalue products [42, Ch. 9, H.1.g.], we have

$$\max_{\boldsymbol{W_q}} \text{tr}(\boldsymbol{C W_q}) = \sum_{i=1}^{N} \lambda_{C,i} \cdot \lambda_{q,i} \tag{27}$$

where the eigenvalues of $\boldsymbol{C}$ and $\boldsymbol{W_q}$ are in the same descending order, $\lambda_{C,1} \geq, \ldots, \geq \lambda_{C,N}$, and $\lambda_{q,1} \geq, \ldots, \geq \lambda_{q,N}$, and the sum of their eigenvalue products yields the maximum value for $\text{tr}(\boldsymbol{C W_q})$ in (27). This maximum value is achieved if and only if the eigenvectors of $\boldsymbol{C}$ and $\boldsymbol{W_q}$ align, that is, $\boldsymbol{U_q} = \boldsymbol{U_C}$, where the eigenvectors $\boldsymbol{U_C}$ are obtained based on the descending order of the corresponding eigenvalues in $\boldsymbol{\Lambda_C} = \text{diag}(\boldsymbol{\lambda_C})$. □

C. Appendix C - Proof For Theorem 3

In the eigenvalue decomposition of $\boldsymbol{W_q^\star}$ as $\boldsymbol{W_q} = \boldsymbol{U_q \Lambda_q U_q^*}$, the diagonal matrix $\boldsymbol{\Lambda_q} \in \mathbb{R}^{N\times N}$ contains the eigenvalues which signify the beam power allocation. Based on Theorem (2), constraint (14b) can be rewritten under the optimal beam solutions as

$$\text{tr}(\boldsymbol{h}_i^* \boldsymbol{U_q \Lambda_q U_q^* h}_i) \leq \pi_i \tag{28}$$

where $\pi_i = \frac{e_i}{\xi_i T_c}$ $\forall i = 1...K$. Define the row vector $\boldsymbol{r}_i^* = \boldsymbol{h}_i^* \boldsymbol{U_q} = \boldsymbol{h}_i^* \boldsymbol{U_C}$, then the above equation becomes

$$\text{tr}(\boldsymbol{r_i^* \Lambda_q r_i}) \leq \pi_i \tag{29}$$

Define row vector $\boldsymbol{d_i}^* = \textbf{diag}(\boldsymbol{r_i r_i^*})$ for $i = 1...K$, matrix $\boldsymbol{D} \in \mathbb{R}^{K\times K} = [\boldsymbol{d_1^*}...\boldsymbol{d_K^*}]$, and vector $\boldsymbol{b} \in \mathbb{R}^{K\times 1} = [\pi_1...\pi_K]$, then (28) can be re-written as in constraint (18c) in (P_{BP}). Recall from the proof for Theorem 2, the ordering of elements in $\boldsymbol{\lambda_q}$ needs to be the same as in $\boldsymbol{\lambda_C}$, that is, in descending order, so as to maximize Lagrangian by (27) which leads to constraint (18b) in (P_{BP}). □

REFERENCES

[1] 3GPP, "System architecture for the 5G System (5GS); Stage 2," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501, 2020, version 16.4.0.

[2] ETSI, "ETSI White Paper No. 28: MEC in 5G networks," ETSI, Tech. Rep., 2018.

[3] 3GPP, "Service requirements for cyber-physical control applications in vertical domains; Stage 1," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.104, 2020, version 17.3.0.

[4] ——, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.261, 2020, version 17.3.0.

[5] R. Malik and M. Vu, "Energy-efficient offloading in delay-constrained massive mimo enabled edge network using data partitioning," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[6] 3GPP, "Service requirements for video, imaging and audio for professional applications (VIAPA); Stage 1," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.263, 2020, version 17.1.0.

[7] R. Malik and M. Vu, "Optimizing Throughput in a MIMO System with a Self-sustained Relay and Non-uniform Power Splitting," *IEEE Wireless Communications Letters*, pp. 1–1, 2018.

[8] V. Talla, B. Kellogg, S. Gollakota, and J. R. Smith, "Battery-Free Cellphone," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 25:1–25:20, Jun. 2017. [Online]. Available: http://doi.acm.org/10.1145/3090090

[9] M. A. Abouzied, K. Ravichandran, and E. Sanchez-Sinencio, "A Fully Integrated Reconfigurable Self-Startup RF Energy-Harvesting System With Storage Capability," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 3, pp. 704–719, 2017.

[10] Powercaster, "TX91501 915 MHz Powercaster Transmitter," http://www.powercastco.com/products/powercaster-transmitter/, accessed: 2018-09-22.

[11] Ossia, "Cota: Real Wireless Power," http://www.ossia.com/cota/, accessed: 2018-09-22.

[12] Energous, "Far Field Wattup Transmitter," "http://energous.com/technology/transmitters/#farfield", accessed: 2018-09-22.

[13] Y. Zeng, B. Clerckx, and R. Zhang, "Communications and signals design for wireless power transmission," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2264–2290, 2017.

[14] O. Galinina, H. Tabassum, K. Mikhaylov, S. Andreev, E. Hossain, and Y. Koucheryavy, "On feasibility of 5G-grade dedicated RF charging technology for wireless-powered wearables," *IEEE Wireless Communications*, pp. 28–37, 2016.

[15] X. Hu, K. Wong, and K. Yang, "Wireless Powered Cooperation-Assisted Mobile Edge Computing," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2375–2388, April 2018.

[16] F. Zhou, Y. Wu, H. Sun, and Z. Chu, "UAV-Enabled Mobile Edge Computing: Offloading Optimization and Trajectory Design," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[17] Y. Zhao, V. C. M. Leung, H. Gao, Z. Chen, and H. Ji, "Uplink Resource Allocation in Mobile Edge Computing-Based Heterogeneous Networks with Multi-Band RF Energy Harvesting," in *2018 IEEE ICC*, pp. 1–6.

[18] C. You, K. Huang, and H. Chae, "Energy Efficient Mobile Cloud Computing Powered by Wireless Energy Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1757–1771, May 2016.

[19] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1784–1797, March 2018.

[20] S. Bi and Y. J. Zhang, "Computation Rate Maximization for Wireless Powered Mobile-Edge Computing With Binary Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 4177–4190, June 2018.

[21] C. You, K. Huang, H. Chae, and B. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, March 2017.

[22] S. Kashyap, E. Björnson, and E. G. Larsson, "On the Feasibility of Wireless Energy Transfer Using Massive Antenna Arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3466–3480, May 2016.

[23] G. Amarasuriya, E. G. Larsson, and H. V. Poor, "Wireless Information and Power Transfer in Multiway Massive MIMO Relay Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3837–3855, June 2016.

[24] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Transactions on Signal and Information Processing over Networks*, June 2015.

[25] J. Xu, S. Bi, and R. Zhang, "Multiuser MIMO wireless energy transfer with coexisting opportunistic communication," *IEEE Wireless Communications Letters*, vol. 4, no. 3, pp. 273–276, 2015.

[26] M. Alhawari, B. Mohammad, H. Saleh, and M. Ismail, *Energy Harvesting for Self-Powered Wearable Devices*. Springer, 2018.

[27] ETSI, "Mobile-Edge Computing - Introductory Technical White Paper," Huawei, IBM, Intel, Nokia, DOCOMO, Vodafone, Tech. Rep., 2014.

[28] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. CUP, 2016.

[29] H. Q. Ngo and E. G. Larsson, "No Downlink Pilots Are Needed in TDD Massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2921–2935, May 2017.

[30] R. Morsi, E. Boshkovska, E. Ramadan, D. W. K. Ng, and R. Schober, "On the performance of wireless powered communication with non-linear energy harvesting," in *2017 IEEE 18th International Workshop on SPAWC*.

[31] B. Clerckx, "Wireless information and power transfer: Nonlinearity, waveform design, and rate-energy tradeoff," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 847–862, 2018.

[32] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.

[33] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2015.

[34] S. Mangiante, G. Klas, A. Navon, Z. GuanHua, J. Ran, and M. D. Silva, "VR is on the Edge: How to Deliver 360° Videos in Mobile Networks," in *Proc. of the W'shop on VR and AR Network*, 2017, pp. 30–35.

[35] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Comms Surveys Tuts*, vol. 19, 2017.

[36] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[37] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient Methods," *lecture notes of EE392o, Stanford University, Autumn*, 2003.
[38] R. Tibshirani, "Newton method," *Notes for Convex Optimization: Machine Learning 10-725, UC Berkeley*, vol. 2019, 2019.
[39] R. Tibshirani, "Subgradient Method," *lecture notes 10-725/36-725: Convex Optimization, Spring 2015*, vol. 2015, 2015.
[40] D. Robbins, C. Cholas, M. Brennan, and K. Critchley, "Augmented and Virtual Reality for Service Providers," Immersive Media Business Brief, Intel Corporation, Tech. Rep., 2017, revision 1.0.
[41] B. Lindmark, "Analysis of pattern null-fill in linear arrays," in *2013 7th EuCAP*, 2013, pp. 1457–1461.
[42] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*. Springer, 1979.